Generating Whole Transcriptomic Profiles Using Compressed Sensing and Machine Learning

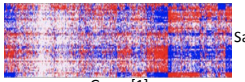
Andra Fehmiu, Kamal Obbad, Dane Hankamer, {afehmiu, kobbad, dhank}@stanford.edu

Motivation

- RNA expression profiles are a critical tool for scientists but can be costly to generate. Here, we show that techniques developed in compressed sensing (CS) can be used to recover high-dimensional RNA expression data from a small number of measurements.
- We use various dictionary learning techniques to represent gene expression in a sparse basis. Our models can infer d-dimensional gene expression data from m-dimensional measurements where m<<d.
- Additionally, we explore using neural nets and other generative models to constrain the solution space (without enforcing sparsity) and reconstruct d-dimensional expression data from m-dimensional measurements.
- Even with a low number number of measurements, we are able to reconstruct gene expression vectors accurately while conserving structure within samples.

Data & Features

Our model is trained on a subsample of data from ARCHS4 which contains full gene expression data across 35K genes and isoforms from 160K sample cell lines. Gene expression is highly structured (shown below) which indicates that CS techniques can be applied. Our subsample consists of 10K samples with 500 genes each.



Pre-processing

- We performed minimal preprocessing to ensure the that our methods were robust to various approaches to measuring gene expression data.
- We standardized our data to so each gene expression falls in the range [0,1].

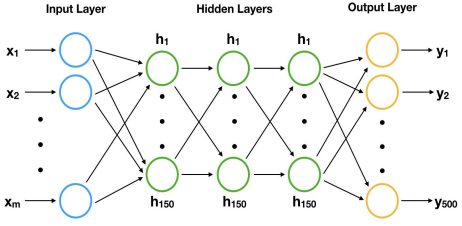
Method & Model

Models:

- **Matrix Factorization Methods:** SVD, k-SVD, sNMF, SMAF
- **MLP Neural Network**

Key Features of Our Model Architecture:

- **Matrix Factorization Sparsity:**
  - k-SVD and sNMF: sparsity constraint imposed by representing k largest active modules in W using the 15 largest eigenvectors
  - SMAF: imposed using Lasso non-negative regularizer to generate U and Orthogonal Matching Pursuit to generate W
  - Conventional SVD: no sparsity constraint imposed
- **MLP Neural Network:**
  - Regularized 3-layered NN with 150 nodes



Model Procedures

- Compute a module dictionary (U) from training data

$$U,\ W = X_{training}$$

- Simulate a random composite measurements Y on the test samples in low-dimensional data.

$$Y = A * (X_{testing} + noise)$$

- Obtain matrix representing module activity levels W using Y and A, the random Gaussian measurement matrix.

$$W_{hat} = Y * inv(A * U) \text{ s.t. } Y \approx A * U * W_{hat}$$

- Recover original high-dimensional gene expression levels.

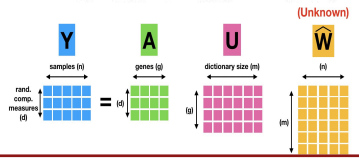
$$X_{testing} = U * W_{hat}$$

SMAF Objective Function

- with OMP constraint to Impose Sparsity Constraint

$$\min_{U,W} \|X - UW\|^2 + \lambda \|U\|_1;$$

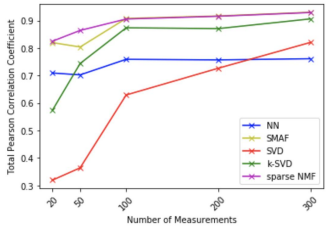
such that $u_{ij} \geq 0, \|u_j\| = 1,$ and $\|w_i\|_0 \leq k\ \forall\ i \in \{1, \ldots, n\}.$



Results

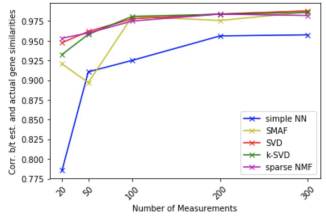
Pearson Correlation vs No. of Measurements

Average pearson correlation between reconstructed gene expression and actual gene expression data exceeds 0.9 for some of our models at higher measurements. SVD performs poorly, as it does not ensure that W has the necessary sparse structure for CS. Sparse NMF, SMAF, and k-SVD perform well, as each algorithm enforces sparsity in W.



Correlation Similarities vs. No. of Measurements

We calculated pairwise pearson correlations between samples and saw that these correlations were highly conserved in reconstructed data. The neural net improved in performance the most as measurements increased, yet it was quite robust for a low number of measurements. Still, matrix factorization techniques outperformed our generative models in maintaining the true distribution of gene expressions across samples.



Conclusion

Main results

- We show that matrix factorization techniques such as SVD, k-SVD, sNMF and SMAF can be used to reconstruct gene expression data with high accuracy.
- We explore whether generative models can be used for accurate reconstruction. We find that MLP neural networks perform relatively well and can even outperform some matrix factorization algorithms with low measurement data.

Future work

- With more computational resources, we can likely increase performance of our neural network. Additionally, we attempted to use GANs and VAEs to represent gene expression data in latent space and then search that latent space to reconstruct full gene expression data from measurements. With more data and CPU resources, we may be able to get these models to perform well.

References

[1] Bora, A., Jalal, A., Price, E., & Dimakis, A. (2017). Compressed Sensing using Generative Models. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17), Doina Precup and Yee Whye Teh (Eds.), Vol. 70. JMLR.org, 537-546.

[2] Cleary, B., Cong, L., Cheung, A., Lander, E. S., & Regev, A. (2017). Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell*, *171*(6), 1424–1436.e18. doi:10.1016/j.cell.2017.10.023.

[3] Dasgupta, S., Gupta, A. (2003). An Elementary Proof of a Theorem of Johnson and Lindenstrauss. Random Struct Algorithms. 22:60–65.